MDPI

Article

Benchmarking Summarization Methods for Scientific Abstracts: From Classical Models to LLMs

Fabio Baumgärtel [1], Enrico Bono [1,2], Paul Perco [1,3] and Matthias Ley [1,2]*

[1] Delta4 GmbH, Vienna, Austria
[2] Division of Pediatric Nephrology and Gastroenterology, Department of Pediatrics and Adolescent Medicine, Comprehensive Center for Pediatrics, Medical University Vienna, Vienna, Austria
[3] Department of Internal Medicine IV, Medical University Innsbruck, Innsbruck, Austria
* Correspondence: matthias.ley@delta4.ai

Abstract

A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: describe briefly the main methods or treatments applied; (3) Results: summarize the article's main findings; (4) Conclusions: indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

Keywords: benchmarking; natural language processing; text summarization; largue language models

1. Introduction

- foo [1]
- bar
  why text summarization in biomedical domain is important
  history of text summarization?

2. Materials and Methods

2.1. Gold-Standard Dataset

To establish a reliable benchmark for automatic summarization, we assembled a gold-standard dataset of 1,000 biomedical articles drawn from a diverse set of peer-reviewed journals hosted on *ScienceDirect* and *Cell Press*. These journals were selected because, in addition to their focus on molecular and biomedical sciences, they provide a standardized *Highlights* section. This section provides concise bullet points that capture the main findings of each article. These served as the reference summaries in our evaluation, while the corresponding abstracts were used as the input texts for the summarization.

Articles were collected systematically across a variety of journals to ensure coverage of different fields within molecular sciences such as drug discovery, genomics, proteomics, biotechnology, and biochemistry. We selected 50 articles from each of the 20 journals, bringing the dataset to 1,000 in total. The distribution of articles across journals is summarized in Table 1.

Received:
Revised:
Accepted:
Published:

Citation: . Title. *Int. J. Mol. Sci.* **2025**, *1*, 0. https://doi.org/

Table 1. Overview of journals and number of articles included in the gold-standard dataset.

Publisher	Journal	Articles included
ScienceDirect	Drug Discovery Today	50
ScienceDirect	Journal of Molecular Biology	50
ScienceDirect	FEBS Letters	50
ScienceDirect	Journal of Biotechnology	50
ScienceDirect	Gene	50
ScienceDirect	Genomics	50
ScienceDirect	Journal of Proteomics	50
ScienceDirect	The International Journal of Biochemistry & Cell Biology	50
ScienceDirect	Cytokine	50
ScienceDirect	Developmental Cell	50
Cell	Cell	50
Cell	Cancer Cell	50
Cell	Cell Chemical Biology	50
Cell	Cell Genomics	50
Cell	Cell Host & Microbe	50
Cell	Cell Metabolism	50
Cell	Cell Reports	50
Cell	Cell Reports Medicine	50
Cell	Cell Stem Cell	50
Cell	Cell Systems	50

This setup provides standardized pairs of abstracts and reference summaries that can be directly used for evaluating automatic summarization methods.

2.2. Summarization Methods

We evaluated 50 summarization methods, ranging from simples frequency-based algorithms to state-of-the-art large language models (LLMs). By having this extensive coverage of methods, we were able to compare established techniques with the latest transformer-based mdodels under identical conditions.

The models were grouped into five categories:

1. Traditional methods: As a foundation for comparison, we included two traditional extractive methods: a simple frequency-based approach and TextRank. These methods provide a simple baseline to compare the more complex approaches with.
2. Encoder-Decoder models: We included a set of pretrained encoder-decoder models, which are available through the HuggingFace library: BART (base and large), T5 (base and large), mT5, a variety of PEGASUS models, and LED. These models are often applied for abstractive summarization and represent well-established neural systems within our benchmark.
3. General-purpose LLMs: <TBD, still have to decide on the final groupings>
4. Reasoning-oriented LLMs: <TBD, still have to decide on the final groupings>
5. Specialized LLMs: To assess whether domain adapatation improves summarization quality, we included large language models additionally trained on medical/biomedical data, such as MedLlama2 and other specialized models.

The complete list of models included in each category is shown in Table 2.

Table 2. Overview of summarization methods/models evaluated in this study, organized by category.

Group	Methods/Models
Traditional methods	textrank; frequency
HuggingFace transformers	bart-large-cnn; bart-base; t5-base; t5-large; mT5_multilingual_XLSum; pegasus-xsum; pegasus-large; pegasus-cnn_dailymail; led_large_16384_arxiv_summarization
Specialized LLMs	medllama2:7b; openbiollm-llama-3:8b_q8_0
General-purpose LLMs	gemma3:1b; gemma3:4b; gemma3:12b; granite3.3:2b; granite3.3:8b; llama3.1:8b; llama3.2:1b; llama3.2:3b; mistral:7b; mistral-nemo:12b; mistral-small3.2:24b; gemma3-tools:4b; phi3:3.8b; phi4:14b; qwen3:4b; qwen3:8b; gpt-3.5-turbo; gpt-4.1; gpt-4.1-mini; gpt-4o; gpt-4o-mini; claude-3.5-haiku; claude-sonnet-4; mistral-medium; mistral-small; mistral-large; gpt-5-nano; gpt-5-mini
Reasoning-oriented LLMs	deepseek-r1:1.5b; deepseek-r1:7b; deepseek-r1:8b; deepseek-r1:14b; gpt-oss:20b; claude-opus-4; gpt-5; magistral-medium; claude-opus-4-1;

With this selection we covered models of different sizes and release periods, ensuring that both widely adopted systems and recent architectures were represented. Extraordinarily large models were not considered as they are impractical for typical summarization pipelines and fall outside the scope of our benchmarking goals.

These 50 diverse models were all tasked with generating summaries for each of the 1,000 abstracts in the dataset, resulting in 50,000 generated summaries up for evaluation.

2.3. Evaluation Metrics

As there is no single metric that can fully reflect summary quality, especially in the biomedical field where both coverage of key information and factual correctness are critical, we used a multitude of metrics grouped into five categories: traditional surface-level metrics, embedding-based similarity metrics, content coverage metrics, factuality metrics and performance-related measures that reflect the feasibility of using the methods in actual real-world applications. By combining all these metrics into one final overall score, we end up with a balanced benchmark value that reflects both summary quality and practical usability.

2.3.1. Surface-level Metrics

This group consists of metrics that compare the generated summaries with the reference summaries mainly at the word or phrase level. While they do not capture meaning beyond surface overlap, they remain common metrics in summarization research and provide a simple foundation for evaluation. We used three ROUGE variants (ROUGE-1, ROUGE-2, ROUGE-L), BLEU and METEOR. ROUGE-1 and ROUGE-2 measure how many unigrams (single words) or bigrams (word pairs) from the reference appear in the generated output, while ROUGE-L identifies the longest sequence of words shared between the two. BLEU calculates how many n-grams in the output also occur in the reference, but it emphasizes precision rather than recall and applies a brevity penalty to counteract the tendency toward overly short summaries. METEOR extends n-gram matching by also considering word stems and synonyms, which makes it more tolerant to variations in wording. Together, these metrics offer a simple but transparent point of reference.

2.3.2. Embedding-based Similarity Metrics

To capture similarity beyond surface-level word overlap, we also included embedding-based metrics built on pretrained transformer models. These methods generate vector representations of the texts, which allows them to capture similarity in meaning rather than just word overlap. Specifically, we employed RoBERTa and DeBERTa, two transformer-based models that have shown strong performance on a variety of natural language processing (NLP) tasks. In the case of summarization evaluation, they can be used to judge whether two summaries capture the same content even if phrased differently.

2.3.3. Content Coverage Metrics

<TBD, maybe also add to Embedding-based Similarity Metrics?>

2.3.4. Factuality Metrics

For factual consistency we used AlignScore, a metric designed to assess whether the statements in a generated summary are supported by the source text. In contrast to the other metrics, we used AlignScore in a way where it does not compare the output to the reference summary but instead aligns it directly with the abstract, as factual correctness can only be judged relative to the input text itself and not against a condensed reference. ... By adding AlignScore we have a metric that is sensitive to errors and halucinations...

2.3.5. Performance Metrics

2.4. Benchmarking Framework

<TBD>

2.5. Computational Resources

<TBD>

- gold-standard data generation
- mention how GenAI was used (e.g. to generate text, data, graphics, assist in analysis, interpretation, ..)
- The use of GenAI for superficial text editing (e.g., grammar, spelling, punctuation, and formatting) does not need to be declared.

This is an example of a quote.

3. Results

This section may be divided by subheadings. It should provide a concise and precise description of the experimental results, their interpretation as well as the experimental conclusions that can be drawn.

3.1. Subsection

3.1.1. Subsubsection

Bulleted lists look like this:

- First bullet;
- Second bullet;
- Third bullet.

Numbered lists can be added as follows:

1. First item;
2. Second item;
3. Third item.

The text continues here.

3.2. Figures, Tables and Schemes

All figures and tables should be cited in the main text as Figure 1, Table 3, etc.

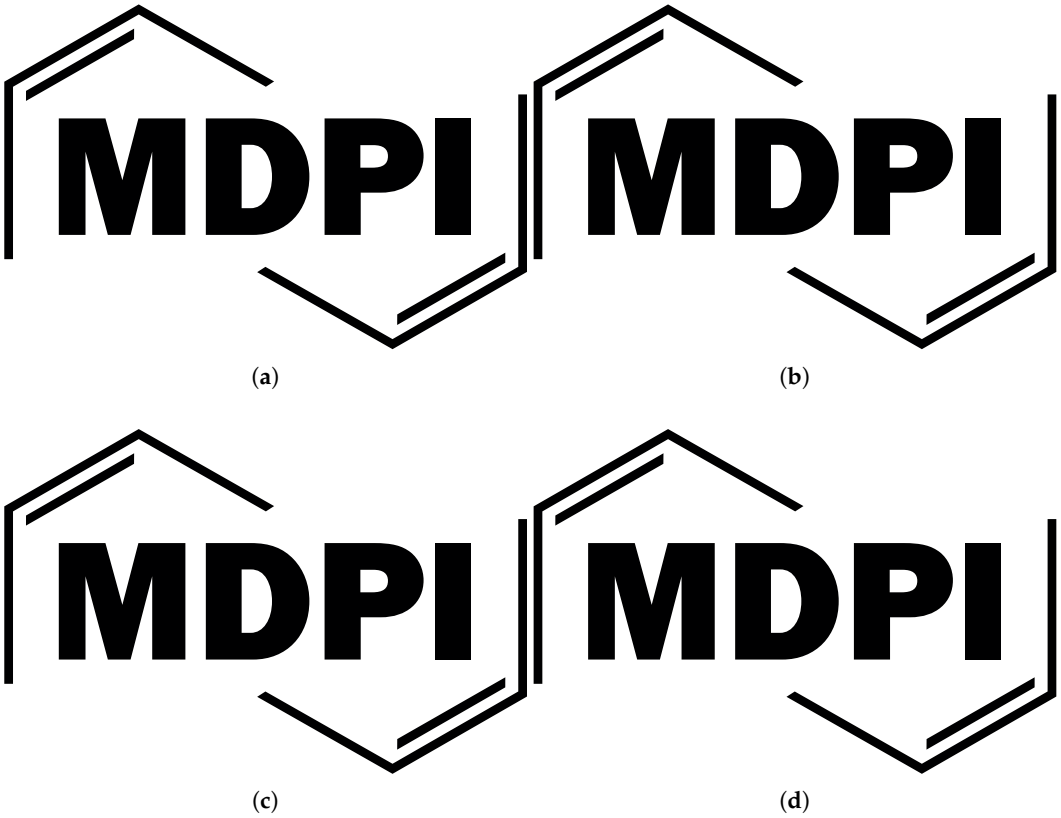
Figure 1. This is a figure. Schemes follow the same formatting.

Table 3. This is a table caption. Tables should be placed in the main text near to the first time they are cited.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data [1]

[1] Tables may have a footer.

The text continues here (Figure 2 and Table 4).



(**a**) (**b**)

(**c**) (**d**)

Figure 2. This is a wide figure. Schemes follow the same formatting. If there are multiple panels, they should be listed as: (**a**) Description of what is contained in the first panel. (**b**) Description of what is contained in the second panel. (**c**) Description of what is contained in the third panel. (**d**) Description of what is contained in the fourth panel. Figures should be placed in the main text near to the first time they are cited. A caption on a single line should be centered.

Table 4. This is a wide table.

Title 1	Title 2	Title 3	Title 4
Entry 1 *	Data Data Data	Data Data Data	Data Data Data
Entry 2	Data Data Data	Data Data Data	Data Data Data

* Tables may have a footer.

Text.

Text.

3.3. Formatting of Mathematical Components

This is the example 1 of equation:

$$a = 1, \tag{1}$$

the text following an equation need not be a new paragraph. Please punctuate equations as regular text.

This is the example 2 of equation:

$$a = b + c + d + e + f + g + h + i + j + k + l + m + n + o + p + q + r + s + t + u + v + w + x + y + z \tag{2}$$

Please punctuate equations as regular text. Theorem-type environments (including propositions, lemmas, corollaries etc.) can be formatted as follows:

Theorem 1. *Example text of a theorem.*

The text continues here. Proofs must be formatted as follows:

Proof of Theorem 1. Text of the proof. Note that the phrase "of Theorem 1" is optional if it is clear which theorem is being referred to. □

The text continues here.

4. Discussion

Authors should discuss the results and how they can be interpreted from the perspective of previous studies and of the working hypotheses. The findings and their implications should be discussed in the broadest context possible. Future research directions may also be highlighted.

5. Conclusions

This section is not mandatory, but can be added to the manuscript if the discussion is unusually long or complex.

6. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualiza-

tion, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.", please turn to the CRediT taxonomy for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

Funding: Please add: "This research received no external funding" or "This research was funded by NAME OF FUNDER grant number XXX." and and "The APC was funded by XXX". Check carefully that the details given are accurate and use the standard spelling of funding agency names at https://search.crossref.org/funding, any errors may affect your future funding.

Institutional Review Board Statement: In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add "The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving humans. OR "The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval)." for studies involving animals. OR "Ethical review and approval were waived for this study due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans or animals.

Informed Consent Statement: Any research article describing a study involving humans should contain this statement. Please add "Informed consent was obtained from all subjects involved in the study." OR "Patient consent was waived due to REASON (please provide a detailed justification)." OR "Not applicable" for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state "Written informed consent has been obtained from the patient(s) to publish this paper" if applicable.

Data Availability Statement: We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section "MDPI Research Data Policies" at https://www.mdpi.com/ethics.

Acknowledgments: In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments). Where GenAI has been used for purposes such as generating text, data, or graphics, or for study design, data collection, analysis, or interpretation of data, please add "During the preparation of this manuscript/study, the author(s) used [tool name, version information] for the purposes of [description of use]. The authors have reviewed and edited the output and take full responsibility for the content of this publication."

Conflicts of Interest: Declare conflicts of interest or state "The authors declare no conflicts of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

Abbreviations

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

Appendix A

Appendix A.1

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

Table A1. This is a table caption.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data

Appendix B

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with "A"—e.g., Figure A1, Figure A2, etc.

References

1. Nguyen, H.; Chen, H.; Pobbathi, L.; Ding, J. A Comparative Study of Quality Evaluation Methods for Text Summarization, 2024, [arXiv:cs/2407.00747]. https://doi.org/10.48550/arXiv.2407.00747.